May 2nd, 1:00 PM - 1:45 PM

Developing and Implementing an Online Research Data Repository for Your University or College Campus

Raymond J. Uzwyshyn
Texas State University, ruzwyshyn@txstate.edu

Developing and Implementing an Online Research Data Repository for Your University or College Campus

Presenter 1 Title

Developing and Implementing an Online Research Data Repository for Your University or College Campus

Session Type

45-minute concurrent session

Abstract

Data-driven research is becoming increasingly important on university and college campuses. Most US federal and many international granting agencies mandatorily require that researchers applying for public grants possess a data management plan and make their research and data publically available through online access. This presentation overviews online research data repositories and implementation strategies for university and college campus libraries. The presentation pragmatically surveys this newer technology landscape and how organizations can begin to think about and implement an online data research repository. This session will survey the landscape but also makes use of practical example from Texas State University and the Texas Data Repository, a large state consortial data repository customizing and utilizing Harvard's open source Dataverse infrastructure.

Location

KIPJ Theatre

Keywords

Research Data Repository, Data Management, Open Data, Federal Grant Compliance, Dataverse, Texas State University, Open Educational Resources

Developing and Implementing An Online Research Data Repository

for Your University or College Campus

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services,
Texas State University Libraries

Online Data Research Repositories What are They?

- Online Way to Manage a Researcher's Data/Metadata
- Permalinking Strategy for Online Data Citation/Access
- Way to Manage Federal Grant Compliance
- Long Term Data Archiving, Preservation, Sharing Strategy

Why are Data Management Repositories Necessary?

Most major Federal grant agencies require data access as mandatory part of the grant proposal/oversite process. (NIH, NSF, NEH, 2013 USDA)

Wordle of the Final NIH Statement on Sharing Research Data, Mandatory 2003

What makes Data Management Repositories useful?

- Leverage and make available faculty, departmental and institutional research
- Allow publication of negative data (less research replication)

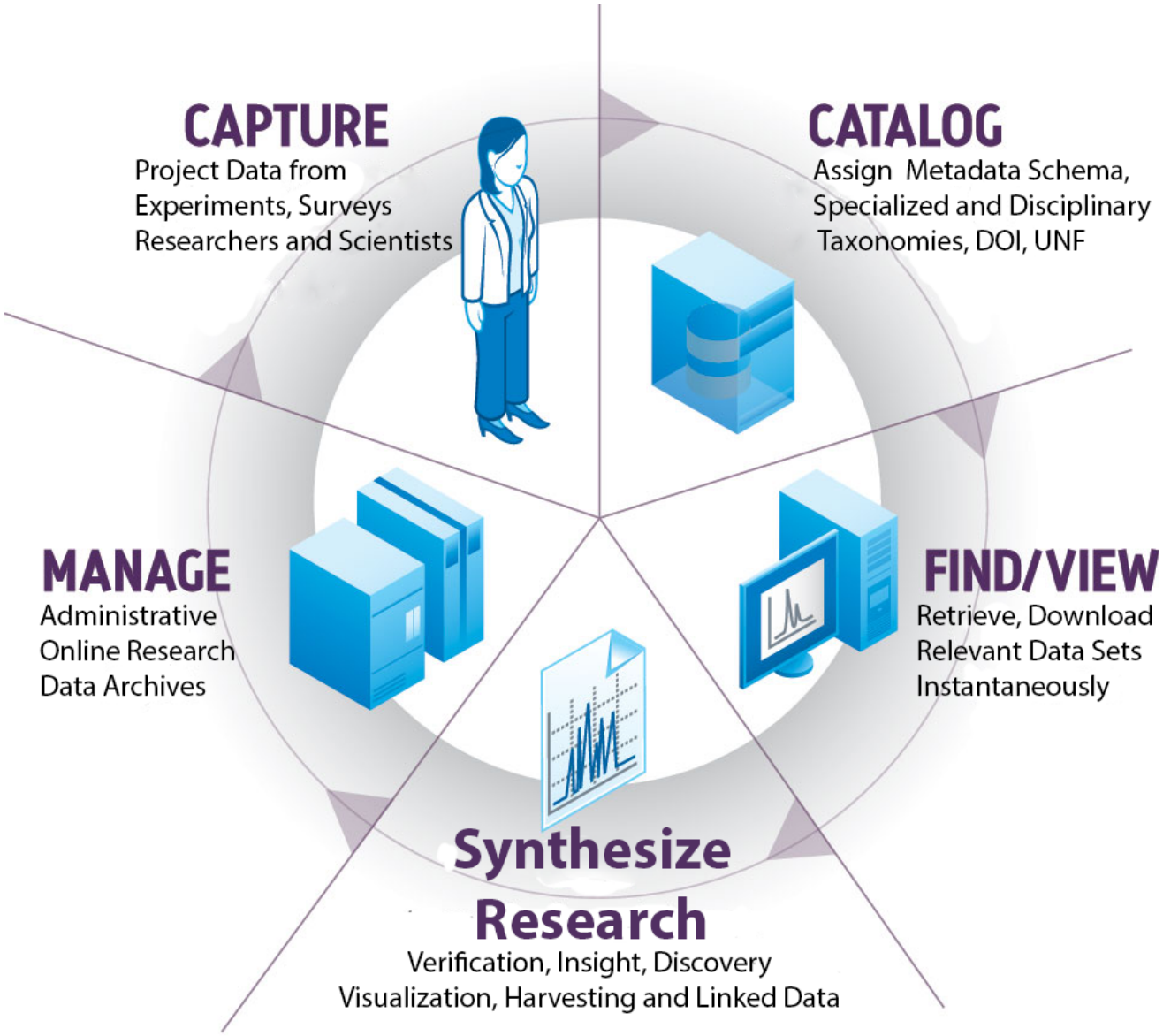
Wordle of the National Science Foundation's Award and Administration Guide*. Chapter VI.D.4, Mandatory 2011*

Data Management Repositories

- Becoming Integral Part of Evolving Science, Social Science and Humanities Research Process (Promote accuracy, efficiency, sharing)

Wordle of the data management policy of the Office of Digital Humanities, National Endowment for the Humanities, 2013

The Research Data Repository Lifecycle



Types of Research Data Repositories

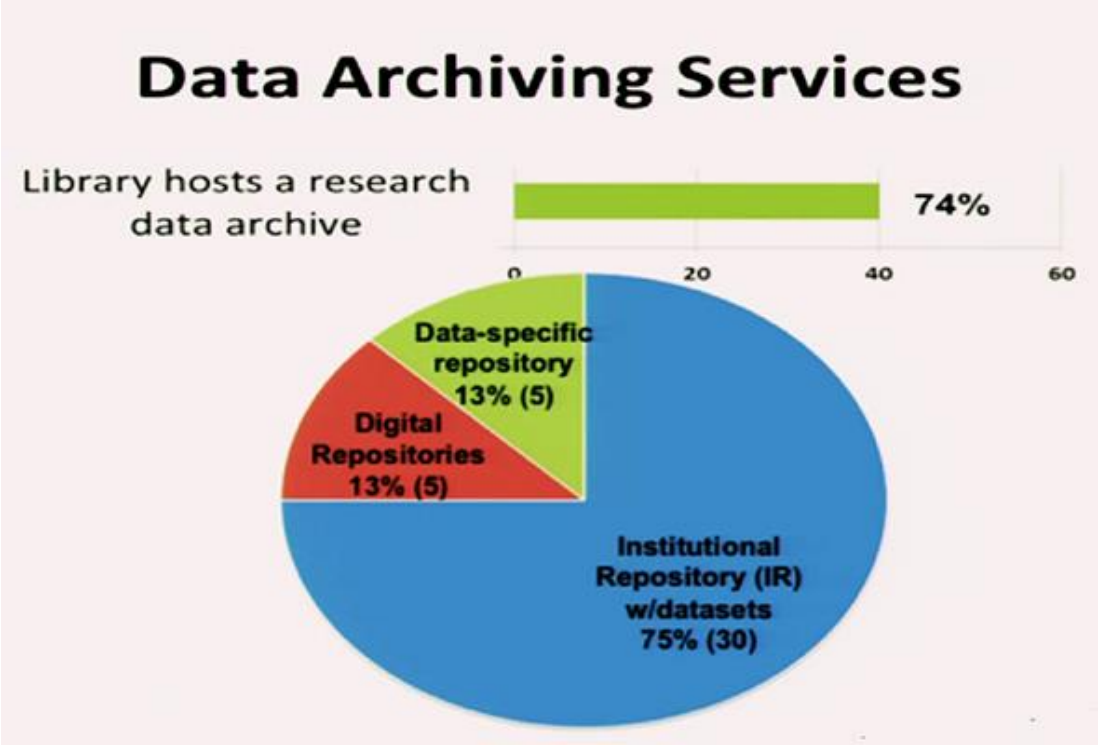
1) Project specific
(usually large single faculty/faculty team projects)

2)Discipline specific
(i.e. Purdue Nanohub/Nanotechnology,
Archeological Data from Academic Center, etc.)

3) Institutional Repository
(either institution wide or consortial)

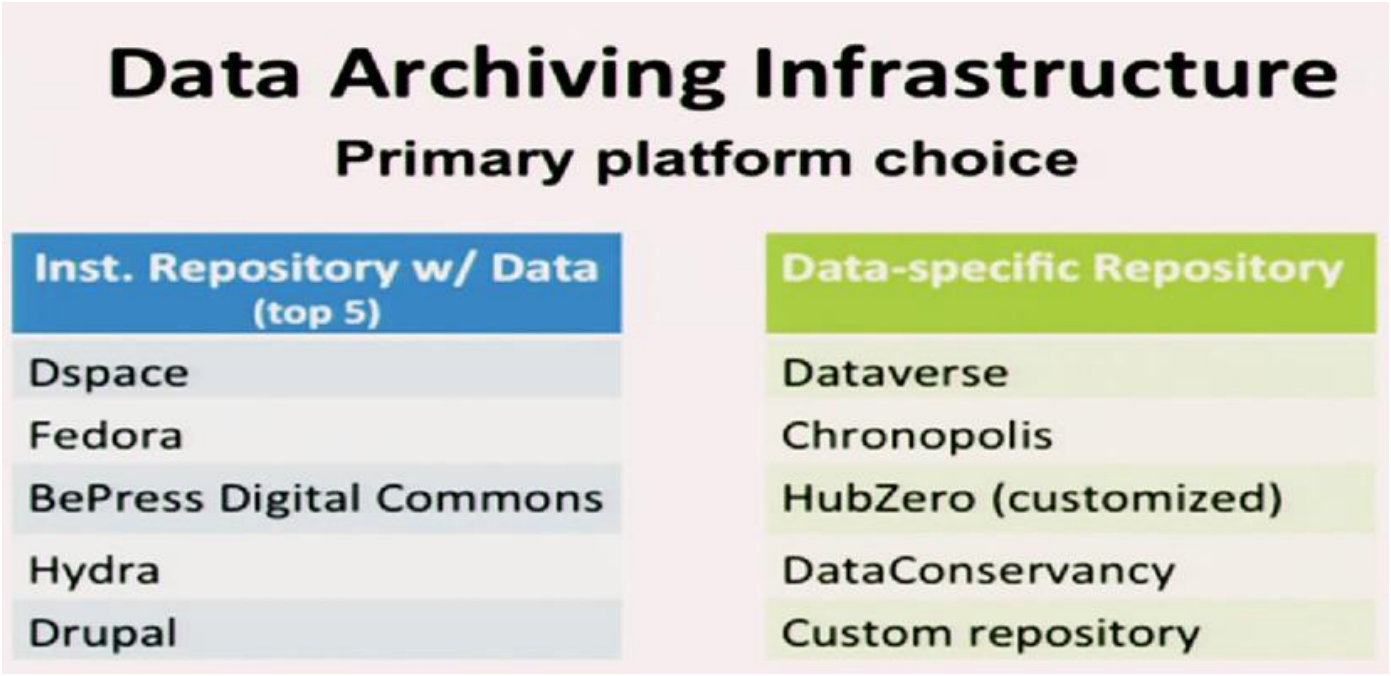
Academic Research Libraries Environmental Scan

Online Data Research Repositories (CNI)



Fearon, D & Sallans, A. C. (January 2014) Institutional Research Data Management: Policies, Planning, Services and Surveys. Coalition for Networked Information. https://www.youtube.com/watch?v=rvbrW7S2fes (54 ARL Libraries currently offer data management services_)

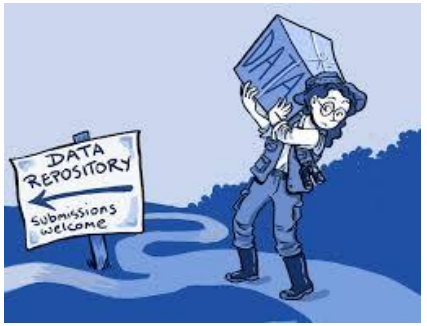
Specific and All-Purpose Data Repository Platforms



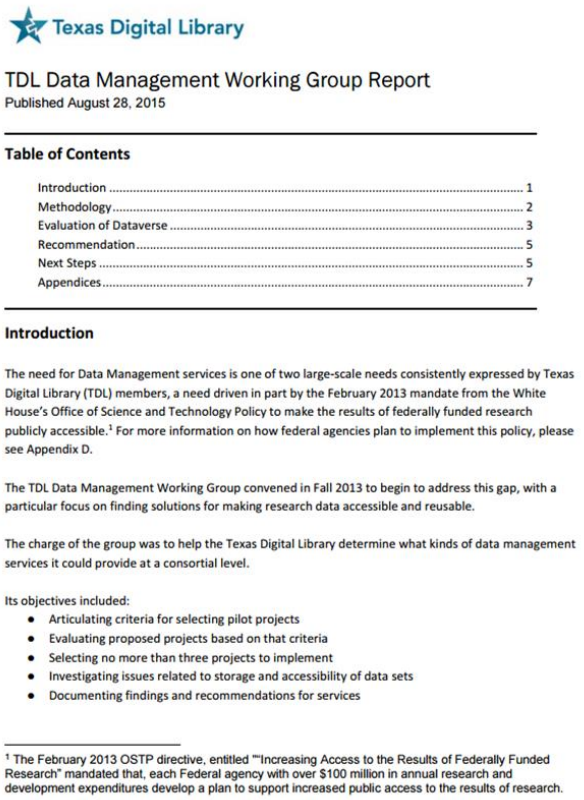
Fearon, D & Sallans, A. C. (January 2014) Institutional Research Data Management: Policies, Planning, Services and Surveys. Coalition for Networked Information. https://www.youtube.com/watch?v=rvbrW7S2fes (54 ARL Libraries currently offer data management services_)

Research Data Repository Software Characteristics

- May be hosted or installed on a university's server
- Each software contains different ranges of management, collaborative options
- Open source and proprietary options
- Ingestion of Various Data Types
  (from Excel to SPSS to more esoteric disciplinary specific formats)



Environmental Scan of Current Possibilities for Your Institution



Texas Digital Library

TDL Data Management Working Group Report

Published August 28, 2015

Table of Contents

Introduction

The need for Data Management services is one of two large-scale needs consistently expressed by Texas Digital Library (TDL) members, a need driven in part by the February 2013 mandate from the White House's Office of Science and Technology Policy to make the results of federally funded research publicly accessible.[1] For more information on how federal agencies plan to implement this policy, please see Appendix D.

The TDL Data Management Working Group convened in Fall 2013 to begin to address this gap, with a particular focus on finding solutions for making research data accessible and reusable.

The charge of the group was to help the Texas Digital Library determine what kinds of data management services it could provide at a consortial level.

Its objectives included:

- Articulating criteria for selecting pilot projects
- Evaluating proposed projects based on that criteria
- Selecting no more than three projects to implement
- Investigating issues related to storage and accessibility of data sets
- Documenting findings and recommendations for services

[1] The February 2013 OSTP directive, entitled ""Increasing Access to the Results of Federally Funded Research" mandated that, each Federal agency with over $100 million in annual research and development expenditures develop a plan to support increased public access to the results of research.

[Data Repository Working Group Report]

(August 28, 2015)

- System Performance
- Robustness
- Usability
- platform availability
- an active open source community

Conclusion: The group recommends that TDL adopt **Harvard's Dataverse** to facilitate the discovery of research data.

The Dataverse Project

Dataverse

Harvard's Open Source Research Data Solution

The Dataverse Network™ Project

A Web Application for Publishing, Citing, Analyzing and Preserving Research Data

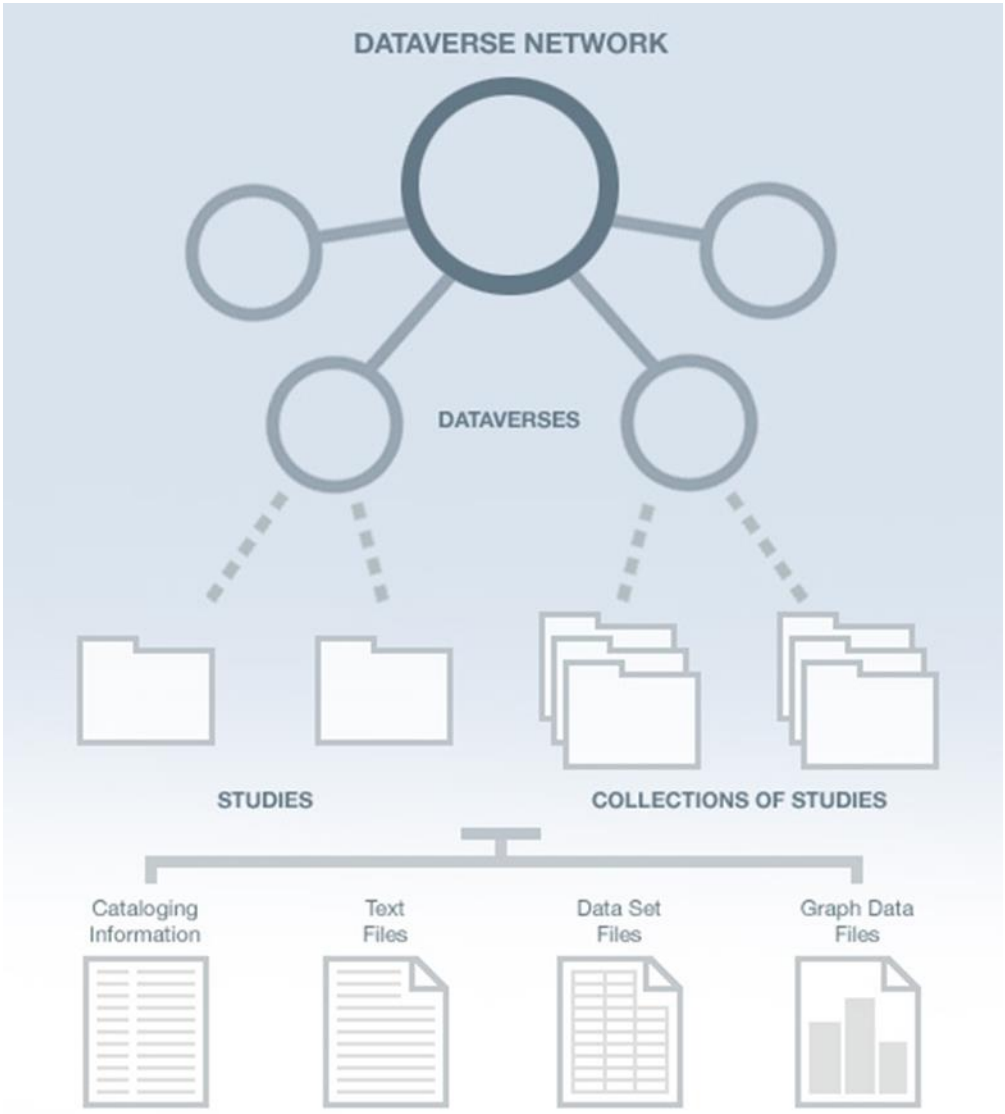
IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

Software framework that enables institutions to host research data repositories

Allows data sharing, control, persistent data citation, data publishing and versioning management

Social Sciences Beginnings (IQSS)
Data Science (site)
http://thedata.org
Dataverse Open Source Download (Github), Software Background

Dataverse Network Architecture



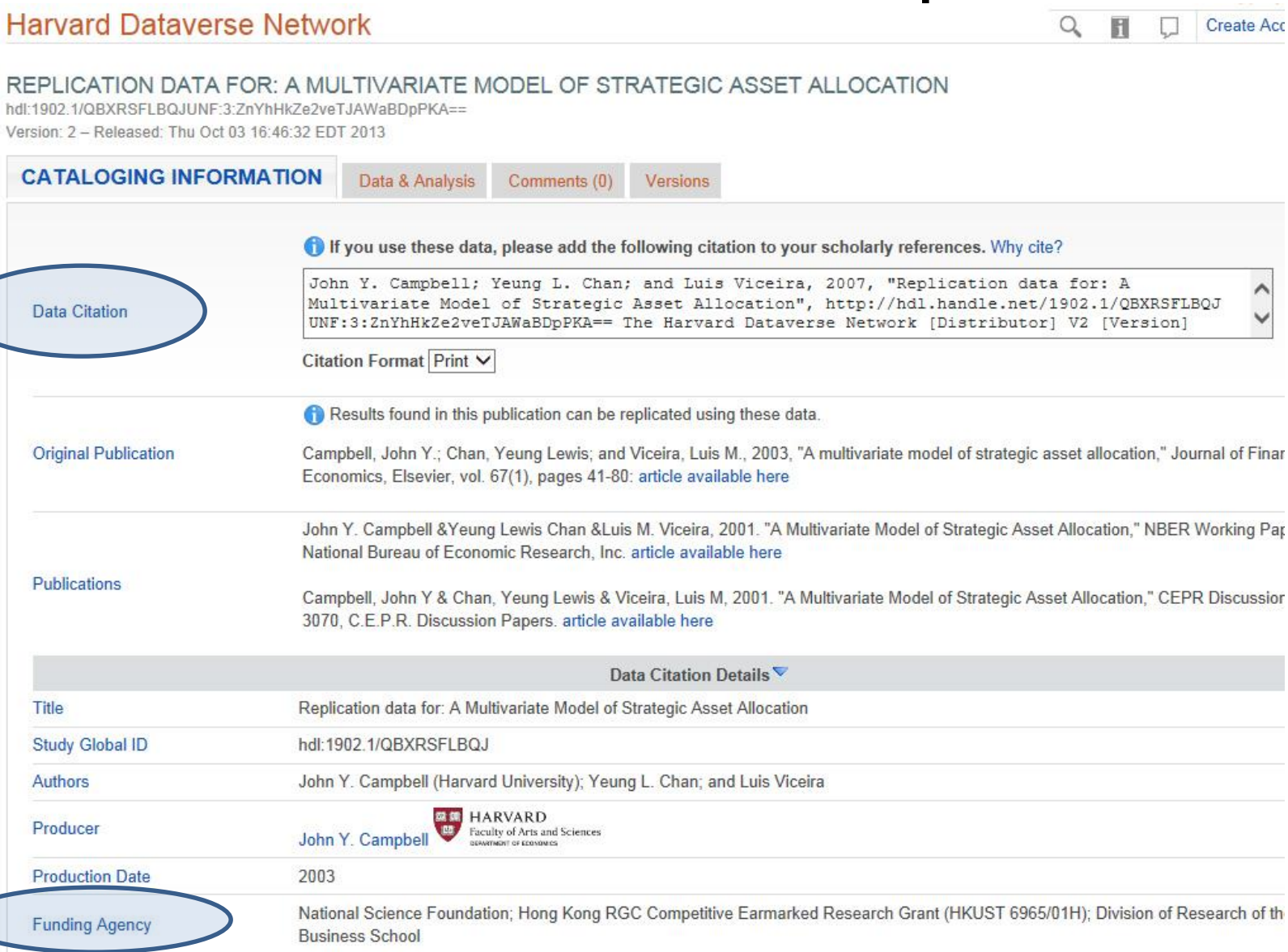
[Why the Dataverse Network? (silent video overview)]

[Open Journal Systems Dataverse Integration]

Research Study Data

Data Set Files
Metadata (Data Describing the data)
Paratextual Research Material (Methodology, Field Notes etc.)
Graph Data Files

Dataverse Data Citation and Metadata Example



Harvard Dataverse Network

Create Acc

REPLICATION DATA FOR: A MULTIVARIATE MODEL OF STRATEGIC ASSET ALLOCATION

hdl:1902.1/QBXRSFLBQJUNF:3:ZnYhHkZe2veTJAWaBDpPKA==

Version: 2 – Released: Thu Oct 03 16:46:32 EDT 2013

CATALOGING INFORMATION | Data & Analysis | Comments (0) | Versions

Data Citation

If you use these data, please add the following citation to your scholarly references. Why cite?

John Y. Campbell; Yeung L. Chan; and Luis Viceira, 2007, "Replication data for: A Multivariate Model of Strategic Asset Allocation", http://hdl.handle.net/1902.1/QBXRSFLBQJ UNF:3:ZnYhHkZe2veTJAWaBDpPKA== The Harvard Dataverse Network [Distributor] V2 [Version]

Citation Format Print

Results found in this publication can be replicated using these data.

Original Publication

Campbell, John Y.; Chan, Yeung Lewis; and Viceira, Luis M., 2003, "A multivariate model of strategic asset allocation," Journal of Finar Economics, Elsevier, vol. 67(1), pages 41-80: article available here

Publications

John Y. Campbell &Yeung Lewis Chan &Luis M. Viceira, 2001. "A Multivariate Model of Strategic Asset Allocation," NBER Working Pap National Bureau of Economic Research, Inc. article available here

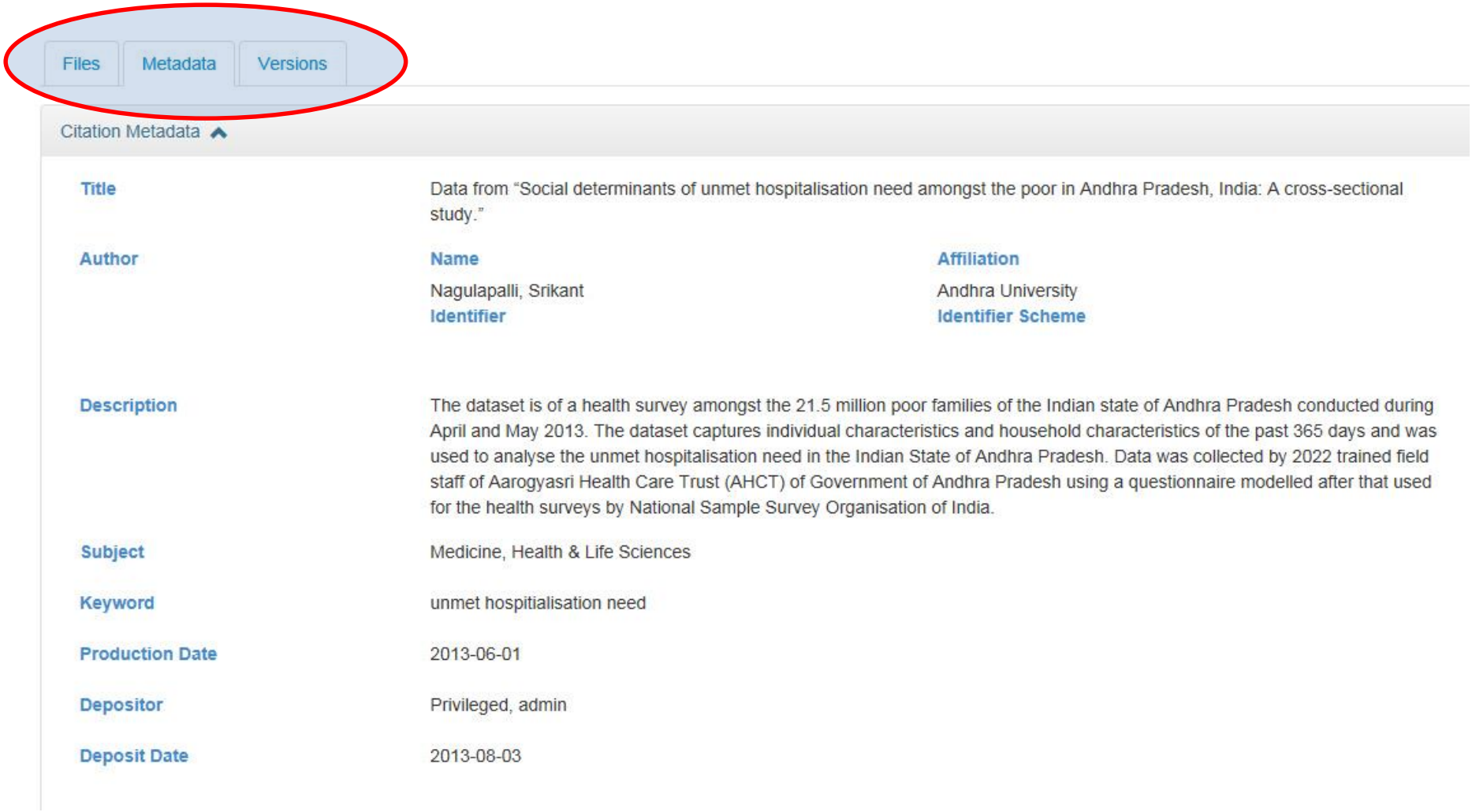
Campbell, John Y & Chan, Yeung Lewis & Viceira, Luis M, 2001. "A Multivariate Model of Strategic Asset Allocation," CEPR Discussior 3070, C.E.P.R. Discussion Papers. article available here

Data Citation Details

Title	Replication data for: A Multivariate Model of Strategic Asset Allocation
Study Global ID	hdl:1902.1/QBXRSFLBQJ
Authors	John Y. Campbell (Harvard University); Yeung L. Chan; and Luis Viceira
Producer	John Y. Campbell HARVARD Faculty of Arts and Sciences
Production Date	2003
Funding Agency	National Science Foundation; Hong Kong RGC Competitive Earmarked Research Grant (HKUST 6965/01H); Division of Research of th Business School

Dataverse Metadata Example

(From the Simple to Complex)



Metadata Schemas Supported: GeoSpatial, Life Sciences, Astronomy and Physics, Georeferenced Data

Data Citation Principles

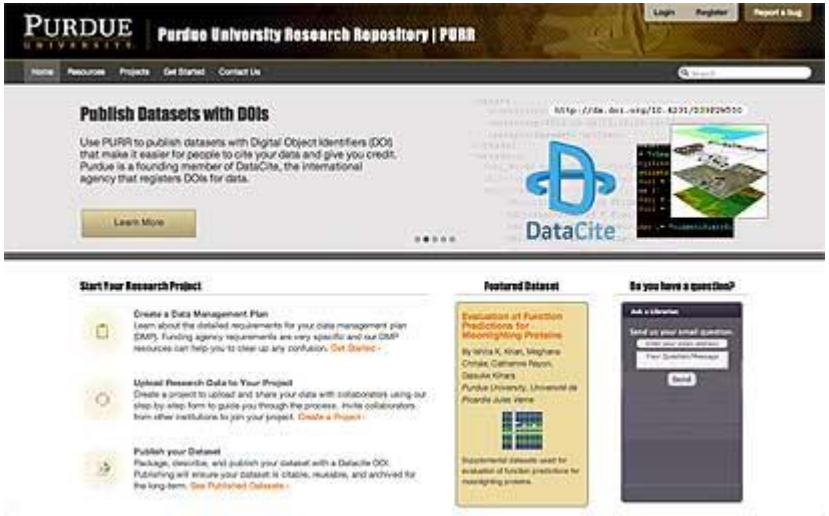
https://www.force11.org/group/joint-declaration-data-citation-principles-final

PURR and Hubzero: Purdue's Data Management System

1.) Create Data Management Plans
2) Collaborate with other Researchers
3) Publish Data Sets (Purdue can publish a DOI: Digital Object Identifier for Data Sets)
Useful For Citation
4) Archive Data Sets

Boilerplate text for data management proposals available

Purr is part of Hubzero platform for scientific collaboration (Originally Nanohub)



- Purr: Purdue University Research Repository (video)
- Purr Site (Proprietary to University)
- Purr Background

Hubzero: Open Source Platform for Scientific Collaboration

Research Collaboration and Data Management Solution

Research Data Types

Spreadsheets
Instrument or Sensor Readings
Software Source Code
Surveys
Interview Transcripts
Images and Audiovisual Files

- https://hubzero.org/
- Getting Started, Downloadable and Hosted Options
- Hubzero Video, Hubzero2

Figshare/Cloud based/Proprietary

Repository where users make their research available in citable, shareable and discoverable manner

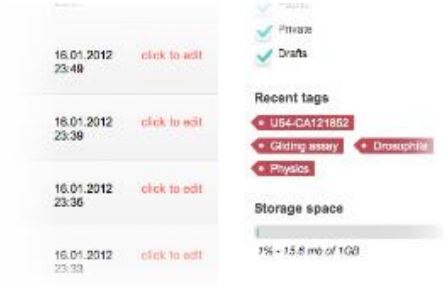
Figures, datasets, media, papers, posters presentations and file sets can be disseminated In a way that the current scholarly publishing Model does not allow

Open Source Platform for Sharing Research

Figshare (video)

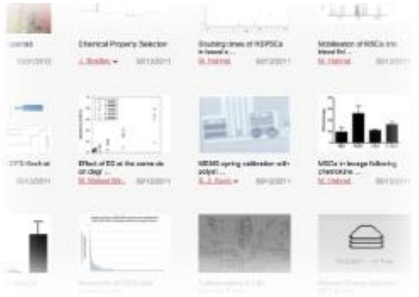
Figshare for Institutions (Video)

Figshare Features (Cloud Based/Proprietary)



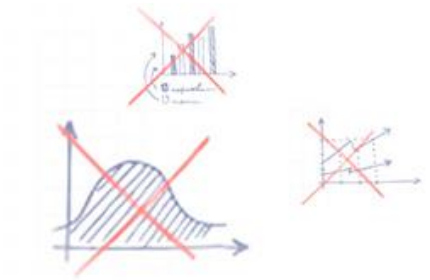
1GB of private space

taggable and easily filtered, your research data is better managed and easy to locate



Unlimited public space

upload to your heart's content
the more - the better



Publish negative data

all published research is citable

Upload all formats

Quick & simple upload

Cloud based

Developing and Testing Your Data Repository

TDL Dataverse Implementation Working Group (August 2015 – December 2016)

Charge: Pilot test, assess, and launch a consortial repository for research data archiving and management.

Main Working Group
& Subcommittees:

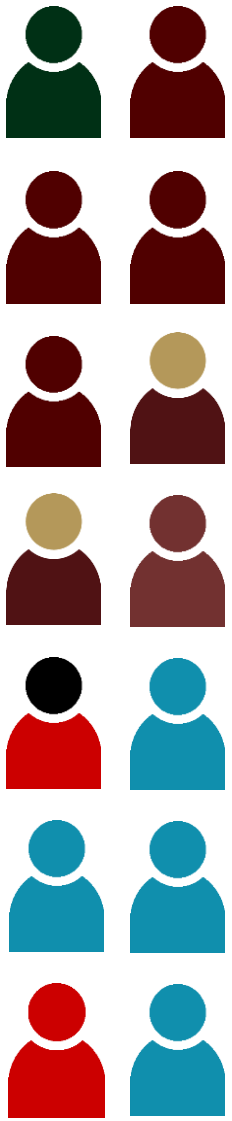
- Policy and Governance
- Workflows and Outreach
- Budget/Business Model
- Technology
- State Data Repository Symposium

[Final Report October, 2016]

14
Working Group members

7
Texas Universities

5
Sub-Committees

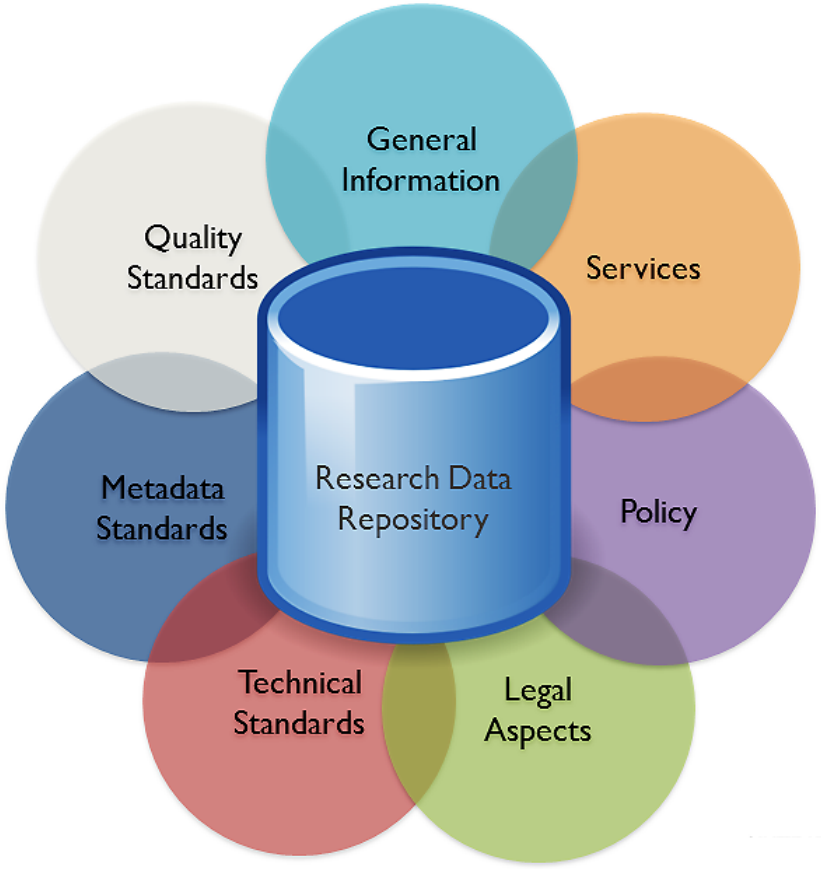


Texas Data Repository

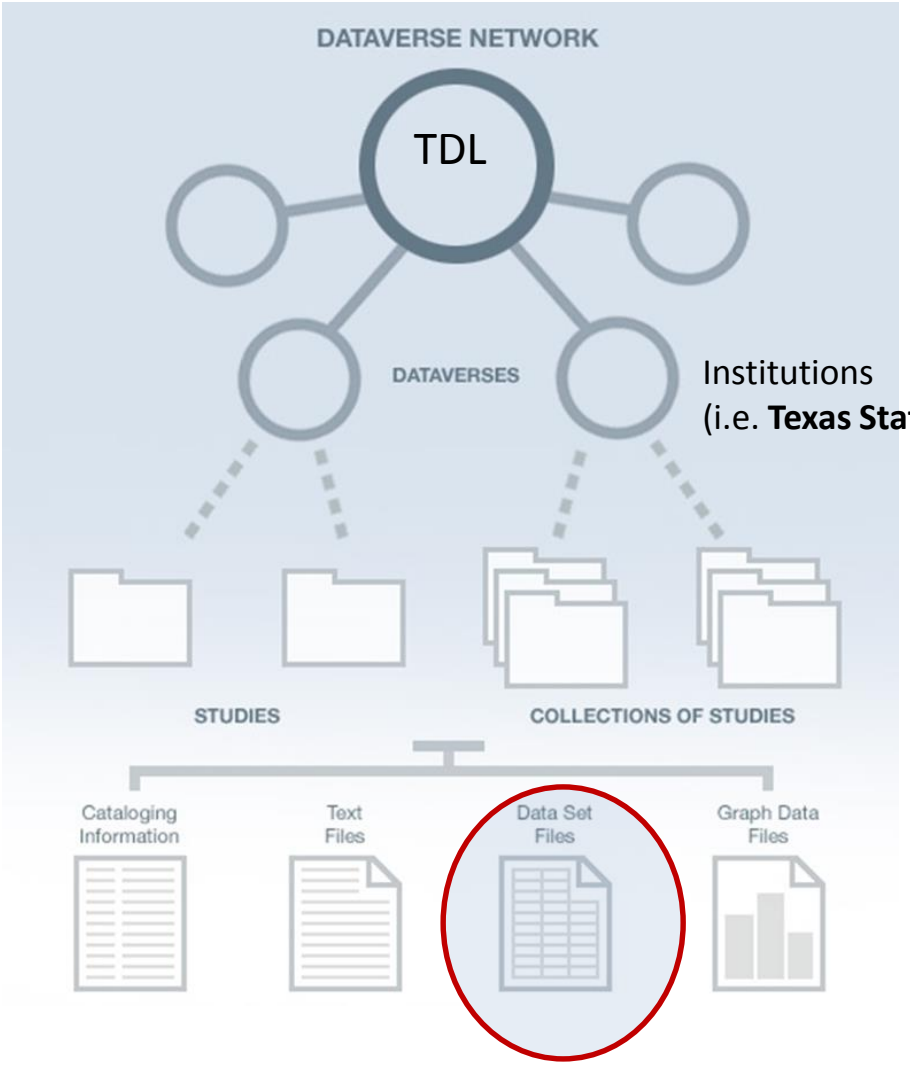
Texas Digital Library Initiative, 2014 -2016

TDL Consortium of 22 universities across Texas leveraging technological cooperation among academic libraries,

The Many Planning Aspects of the New World of Data Research Repositories



Texas Data Repository Consortial Architecture



Institutions
(i.e. **Texas State**)

Research Study Data

Data Set Files
Metadata (Data Describing the data)
Paratextual Research Material
(Methodology, Field Notes , Multimedia, Graphs, Programs etc.)

TEXAS RESEARCH DATA REPOSITORY

Texas Digital Library Test Dataverse

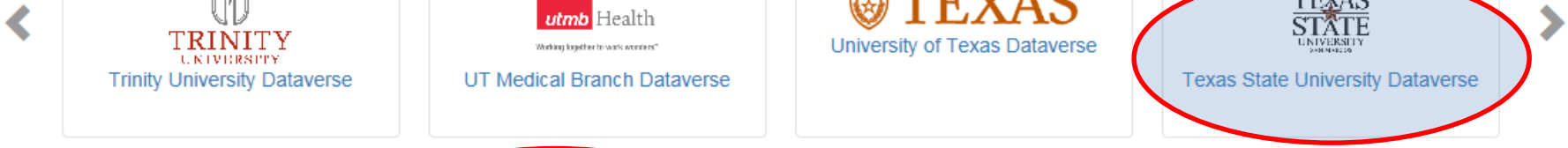
A statewide collaboration of higher education institutions in Texas

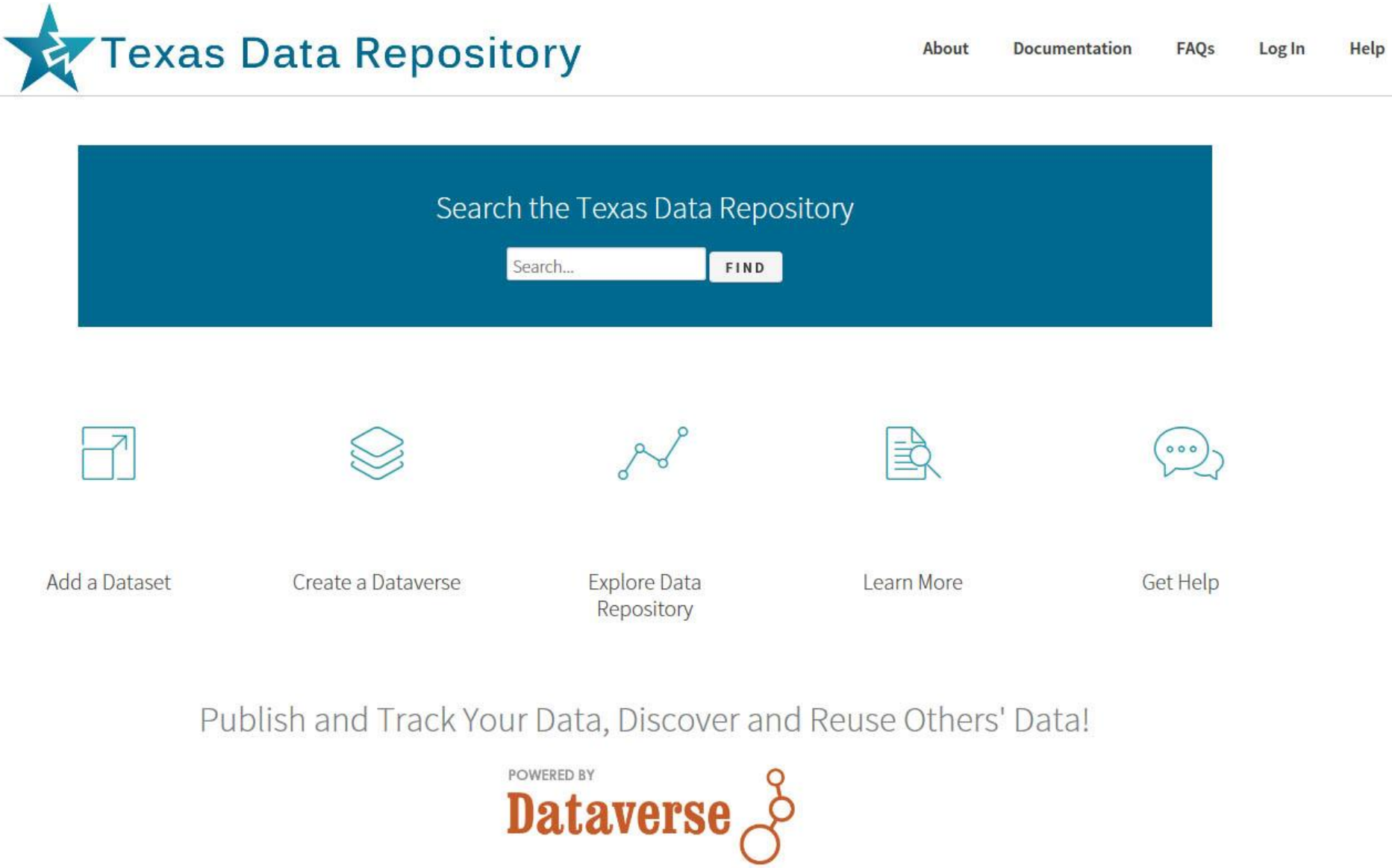
Share, publish, and archive your data. Find and cite data across all research fields.

Welcome to the Texas Digital Library Test Dataverse!

IMPORTANT: This Dataverse server does NOT include the TwoRavens add-on.

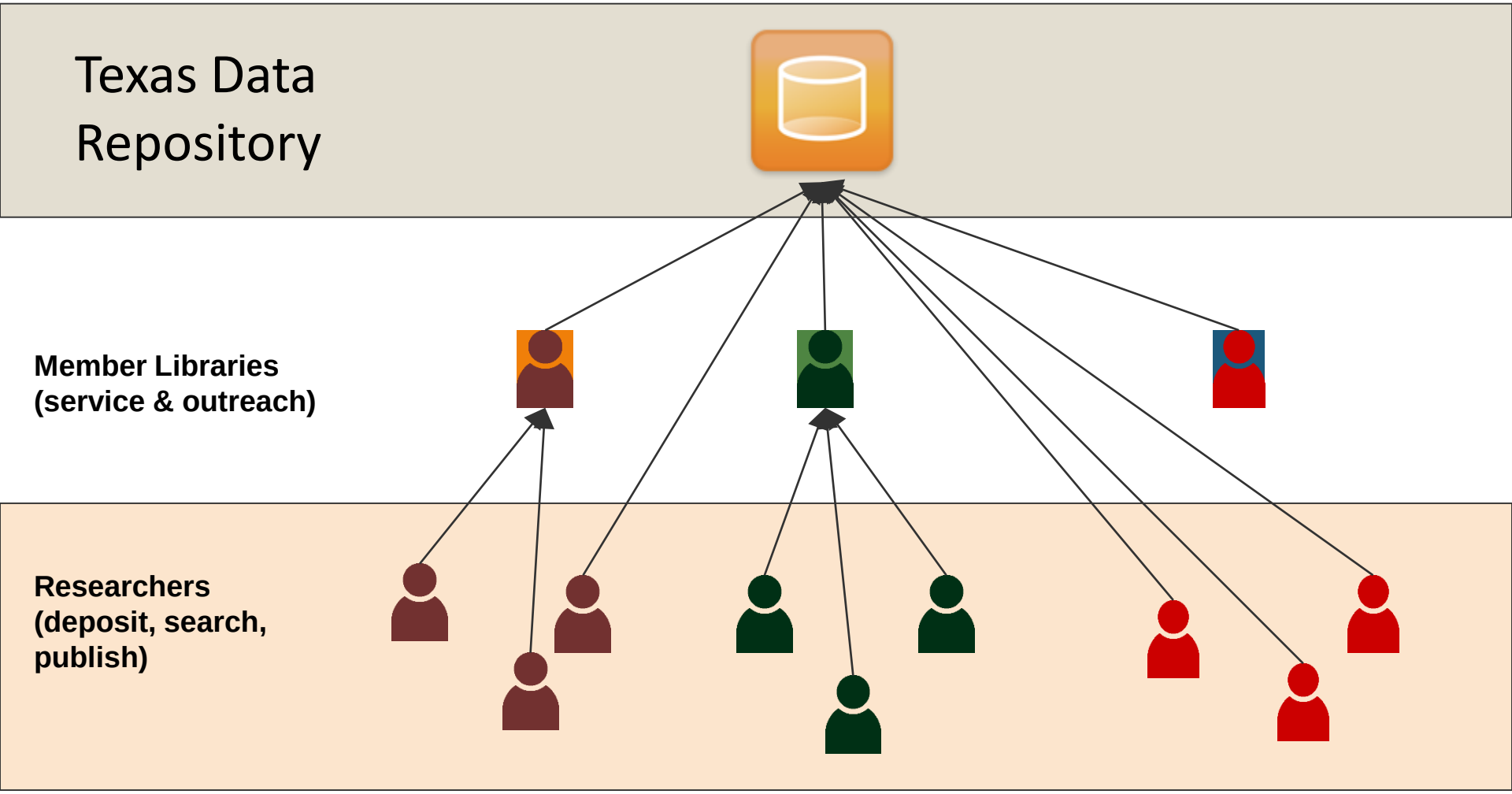
Because of this, you may receive errors when ingesting certain datasets and the "explore" button will not work.





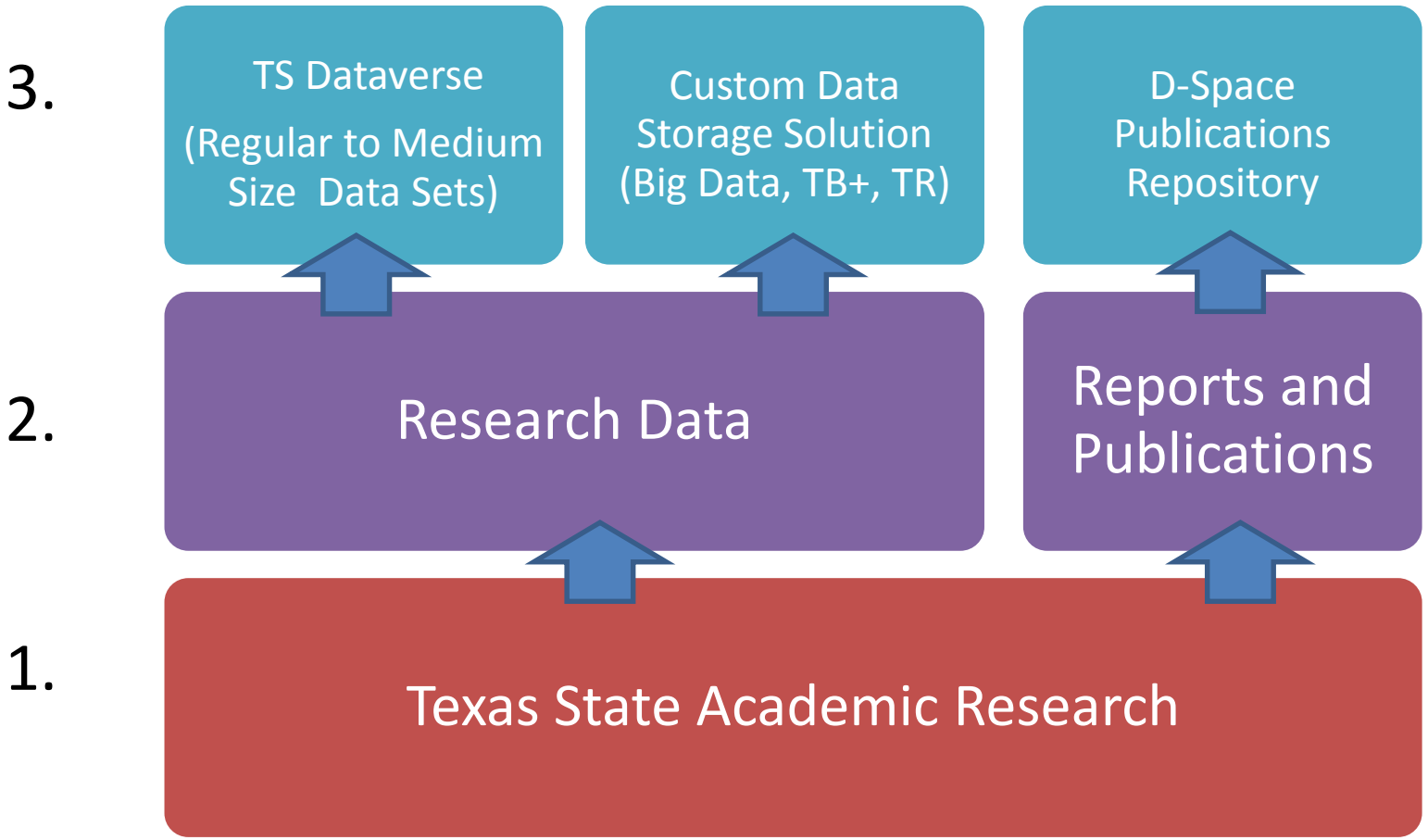
http://data.tdl.org
(UX Usability Focus)

Repository Service Models



Service Models 1) **Mixed** 2) **Mediated** 3) **Unmediated (Direct)**

Texas State Data Repository Architecture



One Size Does Not Fit All
Data Project Needs

Types of Data Projects (Sizes)

1) Normal range

Files/Data Fit on Server/Cloud, may be uploaded, Dataverse, Purr)

2) Large Projects

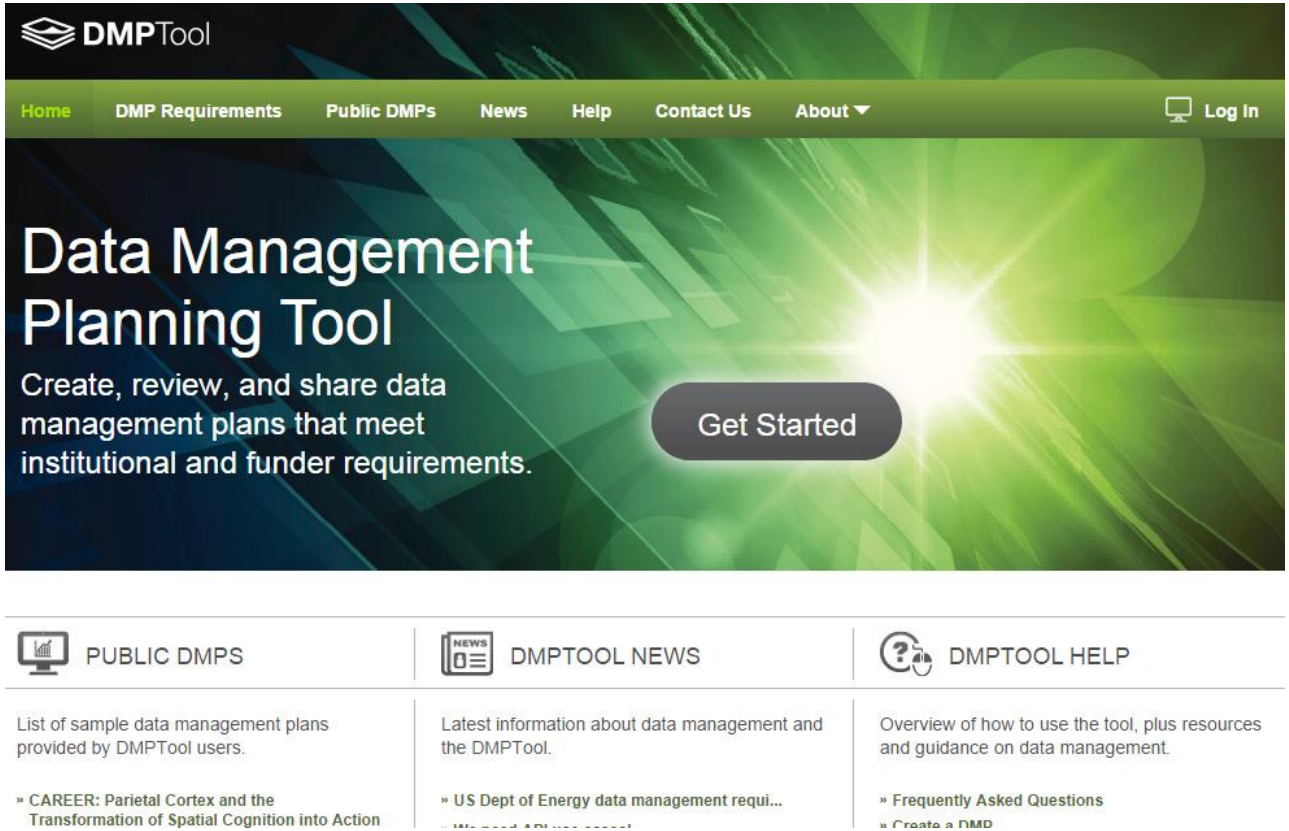
(Data may require specialized university IT Support, i.e. terabyte/petabyte tape drives, Pointers possible)

3) Huge Projects

(Projects require consortial possibilities, national models, Texas Advanced Computer Center TAAC, DEEPN, Duracloud, AWS, Custom Solutions)

Data Management Plan

Documentation/Policy Tool



[Overview Video]

Customizable
Plan Outline Tool
Resource Links
Supports All
Major Funders

Connections with
Office of Sponsored
Research and
Other Relevant
University Offices

https://dmptool.org/
California Digital Library

Data Management Plan Support

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

The Library Supports:

Publication repositories
D-Space

Data repositories, Texas Data Repository

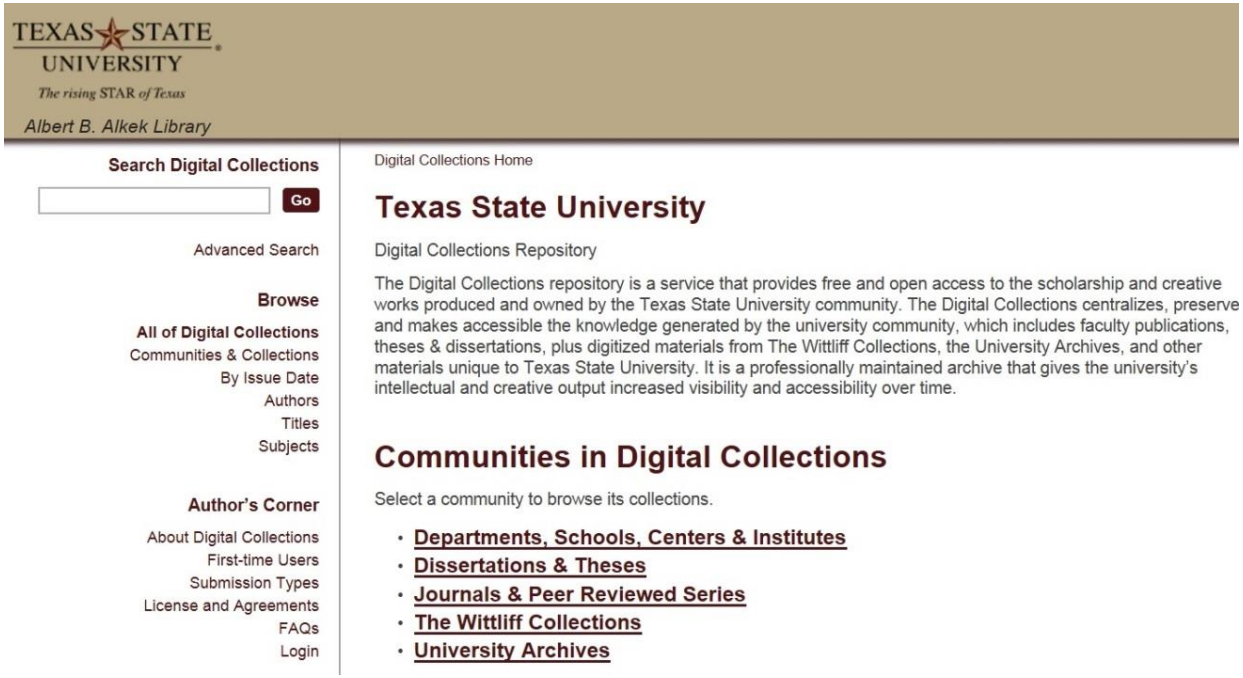
Human Resource Infrastructure

Data Repository Liaison
Publication Repository Liaison
Specialized Metadata Liaison
Subject Liaisons (Outreach)
Committee for Workflows, Standard & Policies

Current/Future Hires
Data Visualization and Analytics Librarian (Tableau, Bayesia)
Digital Collections Librarian (TDR Dataverse/D-Space

Institutional Repository Connections

(MIT, D-Space)



Faculty publications, white papers, preprints, theses, dissertations, working projects

Larger Idea, Grant Compliance, Enabling Faculty Research Online, Raising Research Visibility,

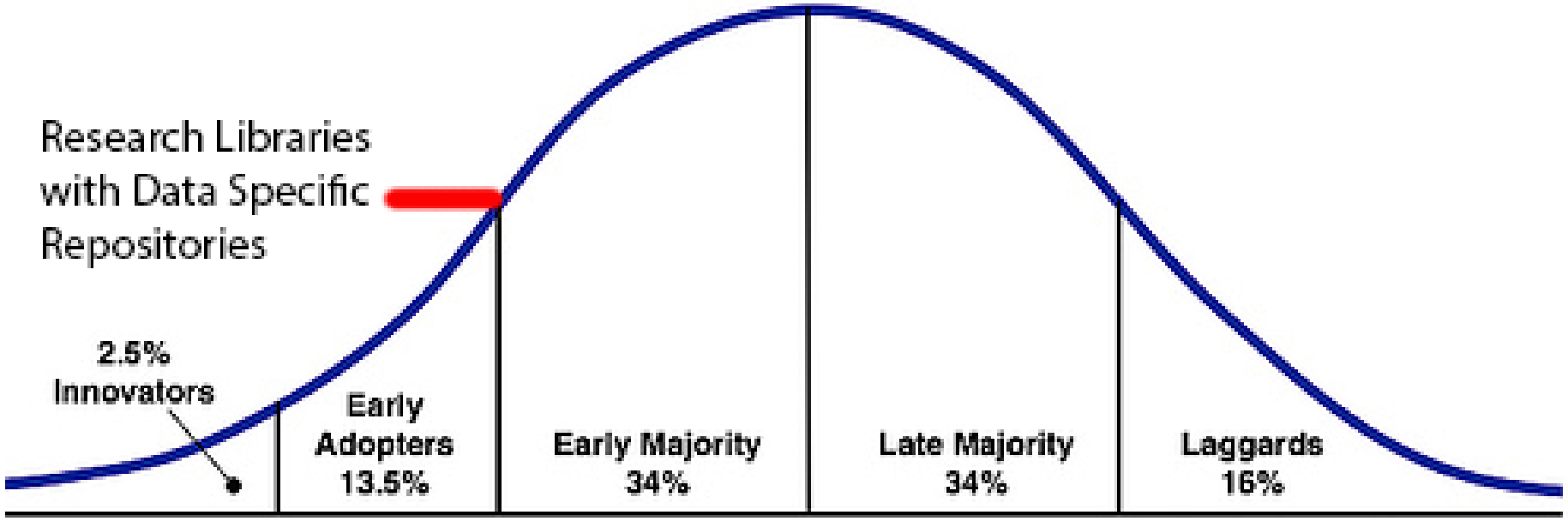
https://digital.library.txstate.edu/

Electronic Thesis and Dissertations (ETD) Repository (D-Space) Possibilities

Co-publish data sets in ETD (D-SPACE) and Data Repository, Links in metadata in D-SPACE and DATA REPOSITORY

Future Possible ETD (D-Space), VIREO, DATA REPOSITORY CONNECTIONS

Data Repository Adoption Lifecycle (2017)



Further Links/References

- ARL NSF Data Sharing Policy and Resource Links, http://www.arl.org/focus-areas/e-research/data-access-management-and-sharing
- ARL (White House Directives and Funded Research Data) http://www.arl.org/focus-areas/public-access-policies#.VoaV0I-cFzo
- Borgman, C. 2015. *Big Data, Little Data, No Data. Scholarship in the Networked Age*. MIT Press
- Baker, Monya. 1500 Scientists Lift the Lid on Reproducibility. www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
- Harris, Richard. (April 2017*). Rigor Mortis How Sloppy Science Creates Worthless Cures*
- California Digital Library DMT Tool: https://dmptool.org/
- Chronopolis: http://www.digitalpreservation.gov/partners/chronopolis.html
- Data Reproducibility Crisis. Nature. http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
- Dataverse. http://thedata.org/
- Dataverse (Data Science Site). http://datascience.iq.harvard.edu/dataverse
- Data Information Literacy Guide. http://www.datainfolit.org/dilguide/
- Data Information Literacy Competencies (Purdue). http://blogs.lib.purdue.edu/dil/the-twelve-dil-competencies/
- DPN (Digital Preservation Network) http://www.dpn.org/
- Duracloud: http://www.duracloud.org/
- Force 11. Data Citation Principles. https://www.force11.org/group/joint-declaration-data-citation-principles-final
- Purr. (Purdue Institutional Data Repository). https://purr.purdue.edu/
- Hubzero. https://hubzero.org/

Further Links/References

- Figshare. http://figshare.com/
- ICPSR Data Management & Curation. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/
- Research Data Management. Principles, Practices, and Prospects (November 2013). *Council on Library and Information Resources*. http://www.clir.org/pubs/reports/pub160
- Cox, A. and Pinfield, S. Research Data Management and Libraries. Jounral of Librarianship and Information Science. June 2013.
- Fearon, D & Sallans, A. C. (January 2014). Institutional Research Data Management: Policies, Planning, Services and Surveys. Coalition for Networked Information. https://www.youtube.com/watch?v=rvbrW7S2fes (video presentation)
- Data Management for Libraries: (LITA Guide) http://www.alastore.ala.org/detail.aspx?ID=10737
- *NMC Horizon Report: 2014 Library Edition.* http://cdn.nmc.org/media/2014-nmc-horizon-report-library-EN.pdf
- "Research Data Management". pp. 6-7 and pp 24 – 45.
- Holden, J. Memorandum for Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Research (2013). http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
- Green, A. Macdonald, S and Rice, R. Policy-making for Research Data in Repositories: A Guide. DISC-UK. http://www.disc-uk.org/docs/guide.pdf
- Research Data Management in the Arts and Humanities (2013). University of Oxford. http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf10-research-data-management-arts-and-humanities (Conference Presentations)
- **Texas Data Repository**. TDR Final Report (October, 2016), Selection Process, Aug. 2015, Peace Williamson et al. UT Arlington, Data Competencies. TDL Texas Data Repository Presentation. Video., Kristy Park, Santi Thompson et al (October, 2016)
- **Uzwyshyn, R**. 2016. Research Data Repositories: The What, When, Why and How of Data Research Repositories *Computers in Libraries*.

Comments/Questions

Pilot Study Responses
Perceived Benefits of Data Repository

- Fulfill federal mandates for sharing publications and research data
- Make research data more widely available
- Statistics on downloads and citations of my data
- Make my data citeable through the assignment of a DOI (digital object identifier)
- Saving various versions of the dataset (data lifecycle)
- Collecting all my data in one place

Collaboration Across Institutions

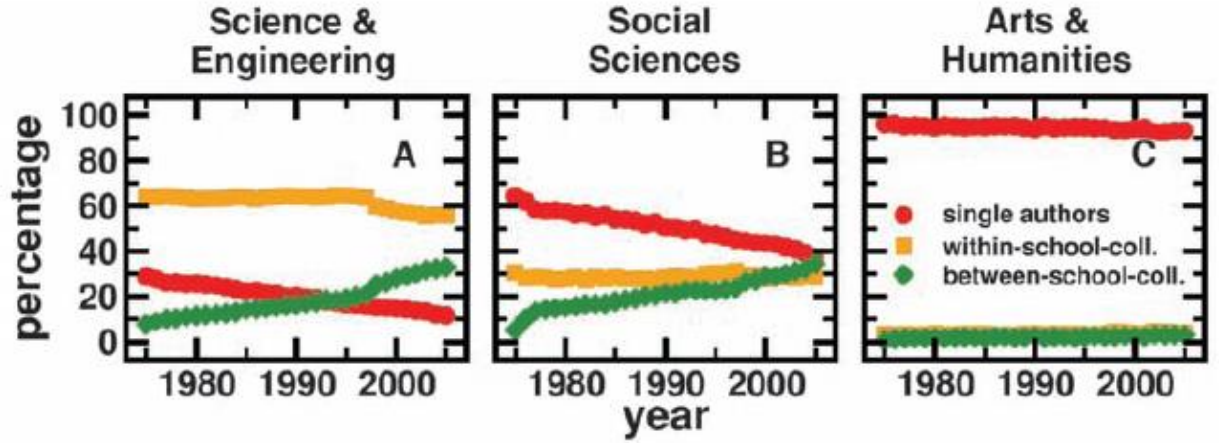


Fig. 1. The rise in multi-university collaboration. By comparing the incidence of papers produced by different authorship structures, we see that the share of multi-university collaborations strongly increases from 1975 to 2005. This rise is especially strong in SE (**A**) and SS (**B**), whereas it appears weakly in AH (**C**), in which collaboration of any kind is rare. The share of single-university collaborations remains roughly constant with time, whereas the share of solo-authored papers strongly declines in SE and SS.

Jones et al. (2008). Science 322: 1259-1262.

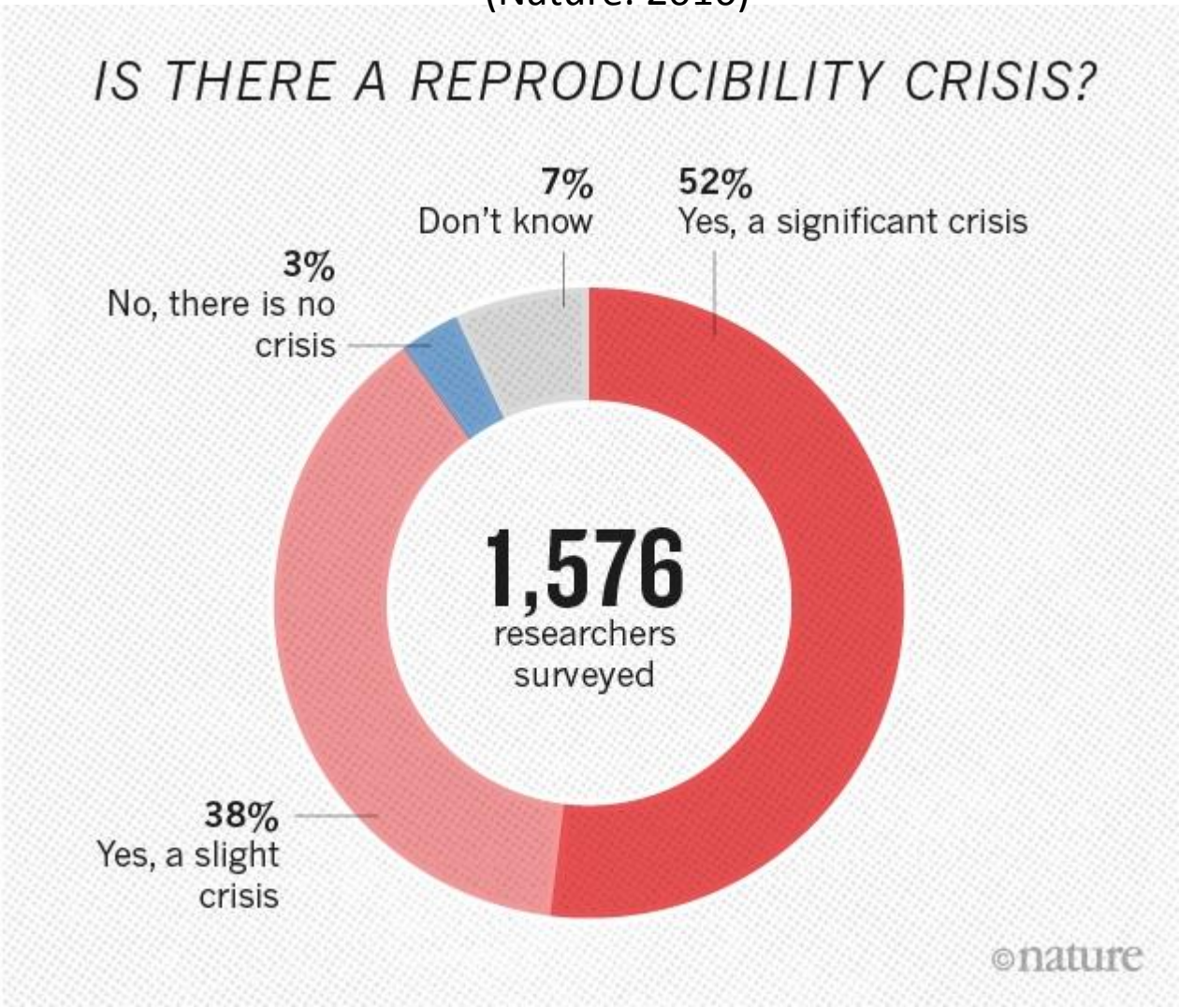
Data Sharing

Currently, 80% of researchers do not share their data

Andreoli-Versbach, P., Mueller-Langer, F. (November 2014). Open access to data: An ideal professed but not practiced. *Research Policy.*, http://dx.doi.org/10.1016/j.respol.2014.04.008

Research Data Reproducibility Crisis

(Nature. 2016)



http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Harris, Richard. (April 2017*). Rigor Mortis How Sloppy Science Creates Worthless Cures*

Hubzero/Purr Customization